学校编码：10384　　分类号_____密级_____
学号：23020051302528　　UDC_____

硕　士　学　位　论　文

聚类融合算法研究及其应用

Research on Clustering Ensemble Algorithms and Their Applications

翁芳菲

指导教师姓名：姜青山教授

专　业　名　称：计算机软件与理论

论文提交日期：2008 年 5 月

论文答辩时间：2008 年　月

学位授予日期：2008 年　月

答辩委员会主席：_____

评　　阅　　人：_____

2008 年 5 月

厦门大学学位论文原创性声明

兹呈交的学位论文，是本人在导师指导下独立完成的研究成果。本人在论文写作中参考的其他个人或集体的研究成果，均在文中以明确方式标明。本人依法享有和承担由此论文产生的权利和责任。

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人完全了解厦门大学有关保留、使用学位论文的规定。厦门大学有权保留并向国家主管部门或其指定机构送交论文的纸质版和电子版，有权将学位论文用于非赢利目的的少量复制并允许论文进入学校图书馆被查阅，有权将学位论文的内容编入有关数据库进行检索，有权将学位论文的标题和摘要汇编出版。保密的学位论文在解密后适用本规定。

本学位论文属于

1.保密（　），在　　年解密后适用本授权书。

2.不保密（√）

（请在以上相应括号内打“√”）

作者签名：　　日期：　　年　　月　　日

导师签名：　　日期：　　年　　月　　日

摘　要

随着信息产业的快速发展，人们迫切需要将大规模数据转换成有用的信息和知识，获得数据间的内在关系和隐含的信息。数据挖掘正是为了解决这一难题而提出的，它结合了统计学、数据库、人工智能、机器学习等技术，并逐渐成为研究的热点。聚类分析是数据挖掘的一个重要研究领域，它是一种无监督的学习方法，通过一定规则将数据按照定义的相似性划分为若干个类，这些类由许多性质相似的数据点构成的，同一个类中的数据彼此相似，与其它类中的数据相异。

聚类融合算法是聚类分析中一个新兴且重要的研究方向。聚类稳定性、准确性和有效性是聚类分析领域中被广泛研究的问题。本文较为系统地分析和研究了聚类融合算法及其在入侵检测方面的应用，藉着分类器组合的思想，提出了一个融合聚类结果的决策机制。首先为了克服传统聚类算法仅在划分某些特定数据集时效果较好的不足和难以确定聚类数的问题，介绍一种基于信息累积的聚类融合算法 EA（Data Clustering Using Evidence Accumulation）。然后针对传统聚类算法和信息累积算法的不足，提出基于模糊 KNN 的聚类融合算法 FNCE（Clustering Ensemble based on the Fuzzy KNN Algorithm），采用对多次运行模糊 KNN 的结果进行融合的方法，累积单次相似信息形成数据间的相似度，从而降低某些不稳定的聚类结果给整个聚类划分结果带来的影响。该方法降低了单一聚类算法受数据分布形状、数据输入顺序、参数变化等因素的影响，提高聚类的准确度，使聚类结果不易陷入局部最优；同时可以根据数据类与类之间的相似度自动确定合适的聚类数，通过实验分析验证了算法的有效性。

入侵检测是计算机安全问题中一个重要的研究课题。由于网络攻击越来越多样化、综合化以及检测环境的多变性，使得用单一聚类方法进行检测受到一定局限，或者不能检测某些未知入侵，或者检测率不高，不能有效检测异常入侵。基于以上聚类融合方法的研究，提出基于聚类融合的异常入侵检测模型 FNIDM（An Intrusion Detection System Based on the FNCE），通过实验分析验证了模型的有效性。

关键词：聚类融合；相似度；入侵检测

Research on Clustering Ensemble Algorithms and Their Applications

Abstract

With development of information technology, it's critical to extract relationships and connotative information from a large amount of data. So Data Mining was proposed to resolve this problem and comprises statistics, database, artificial intelligence, machine learning and so on. Clustering analysis is an important study field in data mining. It has been applied in many applications of data classification and plays a key role in assessing relationships among patterns of data.

In this thesis, Clustering Ensemble is studied systematically in order to improve stability, accuracy and validity of clustering, which are some of extensively studied problems in clustering analysis. Inspired by the work in sensor fusion and classifier combination, a clustering combination approach had been proposed to measure the similarity between patterns. First of all, we introduce algorithm EA (Data Clustering Using Evidence Accumulation). Then in order to overcome the lack of traditional cluster algorithms and EA, the algorithm FNCE (Clustering Ensemble based on the Fuzzy KNN Algorithm) was proposed. We combine the results of multiple fuzzy KNN partitions and make certain instability results with less impact on the entire results of clustering; thereby it avoids a local optimum and improves the accuracy of clustering. Additionally, without specified number of clusters in advance, it can be automatically determined in the process of clustering.

Intrusion detection is an important component of computer network security, while clustering analysis is a common unsupervised anomaly detection method. So after discussing some related topics, an intrusion detection model FNIDM (Intrusion Detection System Based on the FNCE) was designed and was approved that higher Detection Rate and lower False Positive Rate are got in network attacks according to the experiments.

Keywords: Clustering Ensemble, Similarity, Intrusion Detection

目 录

Contents

第一章 绪 论

数据挖掘是为了分析大规模数据，使用智能方法提取数据模式而产生的一门多交叉学科。聚类分析作为数据挖掘的一项主要功能和任务已成为一个非常活跃的研究领域。而聚类融合这个聚类领域的新兴事物正受到越来越多的关注与研究。这里我们将对这些技术的研究现状以及存在的问题等进行阐述，最后对本文研究内容以及本文的结构安排等进行总体概述。

1.1 研究背景及选题意义

近年来，数据挖掘（Data Mining）[1,2]引起了信息产业界的极大关注，其主要原因是存在大量数据，可以广泛使用，目前的数据库系统可以高效地实现数据的录入、查询、统计等功能，但无法发现数据中存在的关系和规则，更无法根据现有的数据预测未来的发展趋势，导致了“数据爆炸但知识贫乏”的现象。人们迫切需要将这些数据转换成有用的信息和知识，获取的信息和知识可以广泛用于各种应用，包括商务管理、生成控制、市场分析、工程设计和科学探索等；于是，人们结合统计学、数据库、人工智能、机器学习等技术，提出数据挖掘来解决这一难题。数据挖掘逐渐成为研究的热点[2,3]。

数据挖掘技术是人们长期对数据库技术进行研究和开发的结果[4]。20 世纪 60 年代，各种数据是存储在磁带和磁盘上，只能提供历史性的、静态的数据信息。70 年代以来，数据管理系统的研究和开发已经从层次和网状数据库系统发展到关系数据库系统（RDBMS）和结构化的查询语言（SQL），可以通过数据库进行查询、访问等获得动态数据信息，联机事务处理（OLTP）将查询看作只读事务，促进了关系技术的发展和广泛地将其作为大量数据的有效存储、检索和管理的主要工具。80 年代后期，发展到数据仓库、多维数据库以及联系分析处理（OLAP），可以在各种层次上提供回溯的、动态的数据信息[1,4]。尽管联系分析处理工具支持多维分析和决策，对于深层次的分析，如数据分类、聚类和预测，仍然需要其它分析工具，进而发展到数据挖掘阶段，数据挖掘使数据库技术进入了一个更高级的阶段，它不仅能对过去的数据进行查询和遍历，而且能够找出过

去数据之间的潜在联系，提供统计性和预测性的信息，从而促进信息的传递[1-3]。

数据挖掘功能用于指定数据挖掘任务中要找的模型类型。数据挖掘任务一般可以分为两类：描述和预测[4]。描述性数据挖掘任务刻画数据库中数据的一般特性。预测性挖掘任务在当前数据上进行推断，以进行预测。数据挖掘能够挖掘多种类型的模式，以适应不同的用户需求或不同的应用，数据挖掘功能如图 1.1 所示：

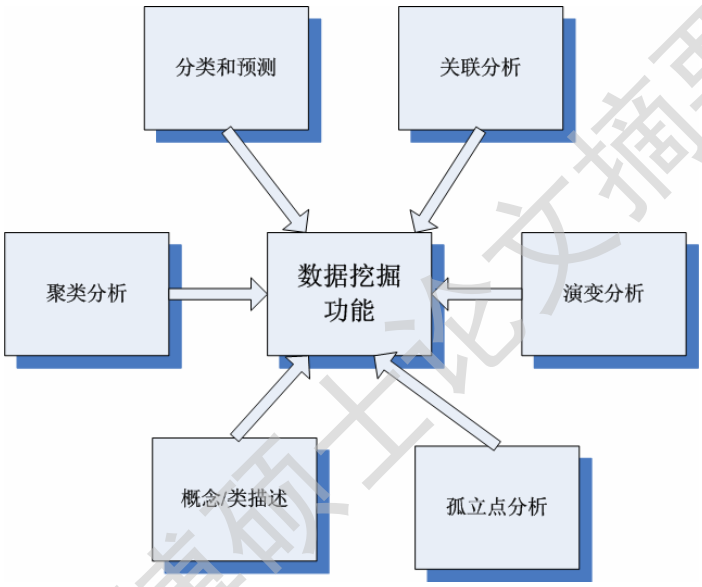

图 1.1 数据挖掘功能

资料来源：数据挖掘概念与技术[1]

聚类分析（Clustering Analysis）[1-4]是数据挖掘的一个基本功能。它是一种无监督的学习方法，其目的是将数据集中的数据人为地划分成若干类，以揭示这些数据的真实分布情况。它通过一定的规则将数据按照定义的相似性划分为若干个类，这些类是由许多在性质上相似的数据点构成的。同一个类中的数据彼此相似，与其它类中的数据相异。目前，人们已经研究出很多聚类分析的方法，聚类分析增强了人们对客观现实的认识，是概念描述和偏差分析的先决条件。概念聚类分析技术的要点[2,4]是，在划分对象时不仅考虑对象之间的距离，还要求划分出的类具有某种内涵描述，从而避免了传统技术的某些片面性。它依据数据对象的特点和对象之间的关系来分组，其目标是使得分在一个组内的对象具有最大的相似性，而分在不同组中的对象具有较大的相异性。

2001年Fred[10]进行了类似聚类融合的研究。聚类融合（Clustering Ensemble）这个词是Strehl和Ghosh[9]于2002年正式提出的，将聚类融合定义为：将多个对一组对象进行划分的不同结果进行合并，而不使用对象原有的特征。其基本思想是，用若干独立的聚类器分别对原始数据进行聚类，然后对这些结果进行组合，最终获得对原始数据的聚类结果。实际数据集形状不规则、有噪声、数据量大，数据是分布式的。而聚类融合因为使用了多个聚类器，可以分布式处理数据，同时噪声和孤立点对结果的影响较小，有良好的稳定性；对不规则数据和噪声数据处理有良好的性能。针对大数据集，采用合适的聚类器，也可以有很好的可伸缩性。聚类融合可以比单一聚类算法得到更好的结果，Topchy总结出了以下几个方面[11]：

1. **鲁棒性：**在各领域和数据集中的平均性能更为优越；
2. **适用性：**在某些数据集上能得到比单一聚类方法更准确的聚类结果；
3. **稳定性：**噪声、孤立点和抽样方法等对聚类结果的影响较小；
4. **并行性和可扩展性：**能对数据子集进行并行聚类并合并；能对分布式数据源或数据属性的聚类结果进行合并。

目前，聚类分析已经广泛地应用到诸多领域中，包括网络数据分析、模式识别、无监督学习、模糊控制、图像处理等方面[1]。其中，在网络应用方面，随着Internet技术的迅速发展，入侵攻击行为也随之增加，如何有效地应用聚类分析进行入侵检测已经成为计算机网络安全问题的一个重要研究课题。

1.2 研究现状及存在问题

聚类分析方法是一种无监督机器学习方法，能作为一个独立的工具来获得数据分布的情况，观察每个类的特点，集中对特定的某些类做进一步的分析。与监督学习方法不同，我们事先对数据集的分布没有任何的了解。迄今为此，人们已经提出了大量的聚类算法[4]，常用的聚类技术可以分为划分方法（Partitioning method）[12-14]，层次方法（Hierarchical method）[8,15,16]，基于密度的方法（Density-based method）[17-19]，基于网格的方法（Grid-based method）[20]以及基于模型的方法（Model-based method）[21,22]。除上述五大类以外，还存在大量的聚类方法，如基于遗传算法的聚类方法[23]，处理高维数据的聚类方法[24]，处理

动态数据的聚类方法[25]，以及将基本聚类方法与各种新技术相结合的聚类方法等[26,27]。

而事实上，任何聚类算法都对数据集本身有一定的预先假设，根据“No Free Lunch”理论[28]，如果数据集本身的分布并不符合预先的假设，则算法的结果将毫无意义，甚至可以说该结果只是给出了一个错误的分布，或是给数据集强加了一个虚构的分布。因此，面对特定的应用问题，如何选择合适的聚类算法是聚类分析研究中的一个重要课题。融合方法将不同算法或者同一算法下使用不同参数得到的结果进行合并，从而得到比单一算法更为优越的结果。在分类算法和回归模型中，融合方法的使用已经比较成熟。但在聚类分析领域，聚类融合方法的研究在近几年才开始出现[29]。

在Strehl和Ghosh提出聚类融合这个概念之前，已经有一些学者对不同聚类结果的合并进行了研究。Fred[10]提出了Co-Association矩阵，用于衡量数据点之间的相似度以及Voting K-means的方法。Strehl和Ghosh[9]提出了三个基于超图的方法: CSPA、HGPA和MCLA。Fred等[30]提出了EA（Evidence Accumulation）方法，用基于MST（Minimum Spanning Tree）的分级聚类算法（Single-Link，Complete-Link，Average-Link）得到最终的聚类结果。Ayad等[31]借用了近邻法的思想，扩展了CSPA的图方法，提出了WSnnG（Weighed Shared nearest neighbors Graph）方法。继而提出了改进的WSnnG方法[32]，文中只使用每个数据点的*k*个最近邻点来生成边，从而得到一个精简的图而减少了计算复杂度。Fern等[33]使用双元图进行聚类融合，提出了HBGF（Hybrid Bipartite Graph Formulation）方法。Topchy等[11]用一个多项式分布的混合模型构建共识函数，然后使用EM算法求解最终的聚类。Ayad等[34]则利用信息论的度量建立一个概率模型，提出了JSDCC（Jensen Shannon Divergence based Clustering Combination）方法。

目前，聚类分析已经广泛地应用到诸多领域中，例如网络数据分析，它是入侵检测[35]的核心问题，通过分析网络数据以检测其中是否包含入侵或异常行为。入侵检测从技术上分为误用入侵检测和异常入侵检测两类。本文主要研究异常入侵检测技术，它通常采用了人工智能的方法，研究者首先选择一种模型，例如神经网络，然后利用这种模型来分析数据。由于缺乏理论上的指导，在选择模型的时候依赖于直觉和专家知识，无法保证其精确性。并且某个模型可能只适用于某

种特殊的数据，因而模型本身无助于我们对根本问题的理解。针对这个问题，Wenke. Lee[36,37]提出了利用信息论的某些概念：熵、条件熵、相对熵和信息增益，可以定量地描述一个数据集的特征，分析数据源的质量，从而为模型的选择提供理论依据。统计分析是最早出现的异常入侵检测技术，IDES，NIDES以及Haystack系统中所包含的异常入侵检测模块都属于这个类别。另一种异常入侵检测技术是基于规则的检测。这种技术所基于的前提与统计异常入侵检测相类似。其区别在于:基于规则的入侵检测系统使用一系列的规则而不是统计出的系统度量来表示系统的使用模式。其代表系统是由数字设备公司（DEC）提出的Time-Based Inductive Machine（TIM）。

在近期入侵检测系统的研究过程中，研究人员提出了一些新的入侵检测技术，这些技术提供了一种有别于传统入侵检测视角的技术层次。例如聚类分析、免疫系统、基因算法的检测等，它们或者提供了更具普遍意义的分析技术，或者提出了新的检测系统架构，因此对异常入侵检测来说，都可以得到很好的应用。其中，Portnoy首先提出了基于聚类分析的异常入侵检测技术[38]，它是一种无监督的异常入侵检测方法，通过对未标识数据进行训练来检测入侵。该方法不需要手工或其它的分类，也不需要进行训练，因此能发现新型的和未知的入侵类型，成为当前的研究热点。

然而使用传统的单一聚类算法进行聚类分析，目前还存在一些不易解决的问题，主要包括以下几个方面[1-4]：

1. **聚类稳定性问题：**聚类算法[2]一般由于初始点选择、数据输入顺序等一些问题，不容易得到稳定而一致的聚类结果；
2. **聚类准确性问题：**一般地，聚类算法很难处理所有形状和大小的数据[4]，甚至对于特定的数据集也很难找出最佳的方法进行聚类分析的问题。至今已经提出了许多方法来改进聚类算法的性能，比如用定义重叠相似度、去除孤立点等方法改进其在类与类之间的重叠数据时的表现，或是用定义密度的方法改进其在类形状非椭球时的表现；
3. **聚类有效性问题：**在大多数实际应用中，聚类数是不可能预先知道的，而且很难预先指定聚类数。现在最常用的办法是反复地使用有效性指标[5-7]对聚类结果进行评估，以此来确定合适的聚类数[3]；

4. **聚类算法可伸缩性问题：**实际应用要求聚类算法能够处理大数据集，且时间复杂度不能太高，最好在多项式时间内完成。目前，已经进行了许多有益的尝试，包括：增量式挖掘、可靠的采样、数据挤压等。比如BIRCH[8]算法中使用CF树就是属于数据挤压技术；
5. **多种数据类型聚类：**现实中的数据对象不仅有数值类型的数据[3]，更多的还包括二元类型、空间数据、多媒体数据、时间序列数据、文本数据、Web数据以及数据流。通常数据对象的属性是由多种类型综合而成。

这些问题不但影响了聚类分析方法对普通数据集聚类分析的性能，而且使得基于聚类分析的异常入侵检测对网络异常数据的分析也存在局限性，例如：聚类的不稳定性可能导致某些网络数据在多次入侵检测中，时而划归为异常，时而划归为正常；聚类的不准确则可能导致某些异常入侵网络数据错误地被划分为正常数据。由此可见这些问题会导致入侵检测系统不能很好地对异常入侵数据进行检测，从而起不到有效的入侵防御作用。

1.3 主要研究内容及特色

本文针对聚类稳定性、聚类准确性和聚类有效性三方面的不足之处，在系统地归纳聚类融合方法的一般原理、方法以及相关技术的基础上，从聚类成员的产生、共识函数的设计等方面，对聚类融合方法进行扩展性研究，并将其应用于入侵检测领域。我们的主要研究内容如下：

1. **研究数据间相似度的度量：**首先介绍了相关的研究工作，然后从聚类成员的产生和共识函数的设计两个方面研究数据间相似度的度量；
2. **研究聚类融合算法：**为了克服传统聚类算法仅在划分某些特定数据集时效果较好的不足和难以确定聚类数的问题，首先介绍一种基于信息累积的聚类融合算法[30]。然后根据传统聚类算法和信息累积算法的不足，提出基于模糊KNN的聚类融合算法FNCE（Clustering Ensemble based on the Fuzzy KNN Algorithm），采用多次运行模糊KNN的方法进行聚类融合，起到了融合结果的作用，不易陷入局部最优及降低参数变化等影响；
3. **研究基于聚类融合的异常入侵检测模型：**在以上聚类融合算法的研究基础上，提出基于 FNCE 的异常入侵检测模型；